MANAGING MISSING DATA IN PATIENT REGISTRIES

Draft White Paper for Fourth Edition of
"AHRQ Registries for Evaluating Patient Outcomes: A User's Guide"

1. Introduction

A patient registry is "an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition or exposure and that serves one or more predetermined scientific, clinical or policy purposes."[1] Registry-based studies, by definition, are observational in nature. Within the context of nonexperimental research, both descriptive and comparative, registries do not typically dictate patient visit schedules, mandate specific diagnostic or laboratory tests, or require patients to complete surveys and patient-reported outcome (PRO) measures at specific intervals. Even during routine care, patients may miss a visit or decline to undergo a procedure or test, and providers may elect to forego expected tests for a few or a specific subset of their patients. Demographics, test results, and other key information may not be documented in the registry due to lack of availability, refusal to provide, or incorrect documentation (e.g., the values are inconsistent or out-of-range). These scenarios, among other potential issues, result in missing data in the registry database. In addition, incorporation of data into a registry from electronic health records, insurance claims, or external observational or experimental studies may introduce missing or disparate data in the registry database.

Missing data can undermine the ability of a registry to make valid inferences both by reducing the study power and, in many cases, by introducing bias. Because missing data can have a substantial impact on registry findings, as well as any clinical or observational study, it is important to take steps throughout the design and operational phases to avoid or minimize missing data. Nonetheless, most patient registries will have at least some missing data. Understanding the types of and reasons for missing data can help guide the selection of the most appropriate analytical strategy for handling the missing data, or the potential bias that may be introduced by such missing data. Once analyses are complete, reports of the registry findings should include information on the amount of missing data and the analytical strategies used to manage the missing data. Interpretation of results must include the potential impact of missing data on the findings, specifically addressing the impact (if any) of the missing data on the ability to infer causality, and should incorporate sensitivity analyses as needed.

The purpose of this paper is to review the types of missing data in patient registries, discuss design and operational strategies for avoiding or minimizing missing data, explore analytic strategies for handling missing data, and consider the impact of missingness on the interpretation and reporting of registry findings. These concepts are discussed in the context of internal validity of studies, or the ability to draw conclusions on the study population, rather than generalizability of results to broader populations who were not enrolled and therefore not represented in the study. The topics discussed are applicable to both retrospective and prospective designs and cover both primary and secondary data sources. Where appropriate, reference is made to other chapters in the document, Registries for Evaluating Patient Outcomes: A User's Guide.[1]

2. Reasons for Missing Data

2.1. Item Nonresponse

Registry data may be missing for many reasons. Item nonresponse, which occurs when a participant completes a case report form (CRF) or survey without providing a response for one or more of the data elements, may be the most common reason. As discussed in the chapter on “Data Collection and Quality Assurance,” CRFs typically incorporate checks to ensure that complete, valid data are entered. These checks may prevent CRFs from being marked as complete if data are missing. However, item nonresponse may still occur, either because CRFs are not marked as complete or because some data elements are optional. As a strategy for reducing the burden of data entry, registries often make only essential fields mandatory for completion of a CRF. The remaining fields are considered optional, and providers may enter only some or perhaps no data into those fields. While these optional fields may not be essential for the primary objective of the registry, they may be critical to support secondary objectives or analyses of subpopulations within the registry. For example, a recent analysis of the characteristics of missing data in three patient registries found that 71% of patients in one registry were missing data for body mass index (BMI), an optional field.[2] Item nonresponse also occurs when patients complete PROs using paper forms and leave some fields blank or enter illegible data.

2.2. Threats from the Left: Truncation

The issue of left truncation, a form of selection bias, arises when events of interest occur prior to a patient’s enrollment in the registry and (typically) pre-empt enrollment in the registry. Applebaum et al. define left truncation as occurring “when subjects who otherwise meet entry criteria do not remain observable for a later start of follow-up.” [3] For example, in a study of miscarriage which enrolls pregnant women, some patients will be left truncated because “an unknown proportion of the source population experiences losses prior to enrollment.”[4] Thus, left truncation results in data missing in the observed cohort due to non-enrollment, leading the study sample to not accurately reflect the underlying target population, in this example, pregnant women at risk for miscarriage.

A related bias can be introduced due to entry of already-exposed individuals into a registry. Consider, for example, a registry designed to study disease progression over several years in patients with a rare disease. Ideally, the registry would enroll only patients at the time of diagnosis, with the goal of collecting detailed baseline and diagnostic information for all patients. However, limiting the registry enrollment to only those newly diagnosed patients would reduce the sample size significantly, and, in the case of a rare disease, likely render the registry infeasible. To enroll sufficient patients, the registry may include both existing (prevalent) patients and newly diagnosed (incident) patients. This enrollment strategy, while practical, has the potential to introduce significant bias for numerous reasons, including under-ascertainment of early events. Examples of the latter include venous thromboembolism risk in women taking third generation OC drugs relative to earlier products, falls after initiating benzodiazepines, and nonsteroidal anti-inflammatory drugs (NSAIDs) and peptic ulcers.[5]

The concept of ‘baseline’ will be different for patients who are newly diagnosed versus those with an existing diagnosis at the time of enrollment, and comparisons of symptoms, treatment effectiveness, and disease progression would need to account for these differences. In particular, the patients with existing diagnoses may be missing information on symptoms at diagnosis or other tests or procedures related to their diagnosis that occurred prior to study enrollment. [6] Ray gives an overview of this issue in the context of medication effects, suggesting that focusing on new users (or newly exposed people, generally) is a strategy which can minimize bias, and should be considered whenever logistically feasible.[5]

2.3. Threats from the Right: Loss to Follow-up, Censoring, Competing Risks

Loss to follow-up and right censoring occur when information is missing at the conclusion rather than the inception of the registry. In studies that collect long-term follow-up data, participants may be lost to follow-up if they formally withdraw from the registry or simply stop completing surveys or coming for scheduled visits. Attrition of this nature occurs for many reasons, including factors both related to the study objectives (e.g., the participant becomes too ill to complete study visits) and unrelated (e.g., the participant moves or changes his/her email address without notifying study staff). Broadly speaking, if the attrition is associated with the study outcomes, it introduces a form of selection bias into the registry that must be described and accounted for in analyses to the extent possible (known as informative censoring in the context of randomized clinical trials). [7] Whether it introduces bias or not, loss to follow-up can limit the ability of the registry to examine long-term outcomes and can have an impact on statistical power. Registries that aim to collect long-term follow-up data are encouraged to develop retention targets, actively monitor retention against those targets, and take proactive measure to minimize loss to follow-up, as needed. Strategies to retain participants and minimize loss to follow-up are discussed extensively in Chapters 3, 5, 10, and 13 of the User's Guide.[1]

A related concept to loss-to-follow-up is administrative right censoring, which occurs when the registry ends before an outcome of interest occurs for all subjects (which is typically the case). This is especially common in pregnancy registries, which are designed to assess outcomes of pregnancies during which the mother (or, in some cases, the father) was exposed to medical products. Pregnancy registries typically collect information on congenital defects that are ascertained at birth or shortly after birth (e.g., 30-day follow-up or, often at most, one year), but are not designed to detect defects or developmental delays that are diagnosed later in life. [8] Right censoring occurs in other types of registries as well. For example, a registry designed to study the effectiveness of a cancer treatment may conduct survival analyses after following patients for five years. Some patients will have died during that period, and their survival after treatment will be known. However, for patients who are still alive at the conclusion of the study, survival after treatment will be right censored due to the close of the registry. In general, missing data due to administrative right censoring will not introduce bias in analysis, but bias is possible if there are strong temporal trends in risk of the outcome.

Finally, competing risks must be considered. A competing risk is an event that prevents the outcome or outcomes of interest not merely from being observed, but from happening in the first place. For example, in a study of incidence of heart attack, death (by any cause besides heart attack) prevents incident heart attack from occurring; in a study of breast cancer, preventive double mastectomy likewise may be considered a competing risk for breast cancer. Competing risks can lead to missing data in certain settings; sometimes a study may be interested in the risk of breast cancer in all individuals – including those who, due to beliefs about their personal risks of developing breast cancer, undergo a mastectomy preemptively. In such a setting, the breast cancer status that these women would have had, had they not gotten a mastectomy, can be regarded as a variety of missing data; in other cases, competing risks do not lead to such clear instances of missing data. See Lau et al. for a more involved discussion of competing risks and missing data, as well as analytic approaches.[9]

3. Approaches to Prevent Missing Data

While analytic methods are available to address missing data issues, the best approach is to avoid missing data to the extent possible through sound design and registry operation. This section discusses strategies for minimizing the likelihood of missing data throughout registry planning and conduct.

3.1. Strategies in Study Design to Minimize Missing Data

The potential for missing data should be considered throughout the planning and design phases of the registry's lifecycle. First, and perhaps most importantly, the registry should focus on data elements that reflect usual care. Data that are collected as part of routine clinical care are far more likely to be captured in the registry, whether they are captured prospectively on registry CRFs or imported into the registry from a secondary data source. For example, Mendelsohn et al. found that baseline demographic information (e.g., age, gender) and diagnosis, treatment, and outcome variables that were collected as part of routine care were missing at very low rates (generally less than five percent).[2]

Second, the use of 'required' or 'mandatory' fields for primary, prospective data collection can greatly reduce the amount of missing data for key variables. However, the number of required fields must be balanced with the burden to sites and patients participating in the registry (burden is a critical component of site and patient retention). When denoting fields as required, focus should be on the data elements that are necessary to answer the primary research question(s) posed by the registry. Information that is being collected for secondary objectives or to support subgroup analyses may be better designated as 'optional'. In addition, CRFs should include response options for 'not applicable,' thus allowing the registry to distinguish between data that are missing because they are not applicable for some patients (e.g., a laboratory test) and those data that are missing for other, possibly unknown reasons. Poorly worded questionnaires with vague directions can also result in unusable or outright missing data.

Registries that incorporate PROs must take into account several additional considerations to minimize the likelihood of missing PRO data. Collection of PRO data is particularly time-sensitive. Typically, the PRO must be completed within a protocol-specified window (e.g., 30 days). If the PRO is not completed, the data should be marked as missing, since completion of missing PRO forms at later visits may introduce recall bias. The time-sensitive nature of PRO collection underscores the need to carefully select both the PROs and the mode of administration. Selection of appropriate PROs that will pertain to the patients' clinical experience is critical. As noted in Chapter 5 of the User's Guide, PROs should collect meaningful information that is necessary to achieve study objectives. The importance of the PRO measure(s) should be explained clearly to the patients. Use of validated PROs can ensure that the surveys are clear, concise, and appropriate for the study population. When possible, use of PROs that are already being completed as part of the standard of care can improve response rates. Mode of administration for the PRO is also critical and should take into account the age, disease severity, and computer literacy of the target population. As an example, a study that intends to enroll college students may see a better response rate with a web-based PRO as opposed to paper forms that must be received and returned by mail. Depending on the resources available and the needs of the target population, offering multiple modes of administration (e.g., paper, telephone, and web-based) and allowing patients to switch between modes may encourage higher levels of completion, although this can also create other risks to data completeness. Missingness can vary slightly based upon the modality, e.g., a web-based or orally administered PRO can prompt or require a respondent to answer a question to proceed unlike a paper-based tool. To this end, different modes of data collection should ideally only be considered when appropriate psychometric evaluation has been performed on the given PRO for each administration method.

Pilot testing of registry CRFs and surveys, including PROs, is highly recommended to uncover issues with clarity, length, or availability of data and therefore minimize missing data. Pilot tests with patients, clinicians or other parties not on the registry planning team can proactively identify questions with unexpectedly high rates of missing data and provide registry developers with the opportunity to explore reasons for the missing data and take corrective action (e.g., provide additional training, revise questions for clarity, remove questions that are not routinely collected).

If existing data sources (e.g., other registries, electronic health records (EHRs), claims databases) will be used to complete specific fields in the registry, due diligence and pilot testing is critical to understand completeness of the variables and potential impact of missing data on the study objectives. Secondary data originate from many different sources and are often collected with non-research motivations (e.g., billing, clinical care). Because of this, these sources may have missing data points, and because the data may be captured through another entity and/or was not collected for research purposes, queries around missing data may not be possible.

EHR data are collected primarily to track clinical care rather than for research purposes. Because these systems are used by a large number of providers and networks under minimal standardized procedures for data recording, they are subject to potentially extensive missing data. In some cases, these data are retrievable through chart review or by locating the data within alternative locations in the medical record; however, the data may not be retrievable due to lack of recording, specificity or standardized fields in a large portion of the patient records. As an example, symptoms of patients who are terminally ill may be poorly documented in an EHR, and this missing information may introduce bias into a study by implying (by omission) that these patients are healthier than they are. [10] Conversely, healthier patients may have fewer health care encounters in less severe conditions. Continuity of care can also be an issue. Many EHRs provide documentation of care provided in either inpatient or outpatient settings, but not both, which can lead to missing data and systematic bias. For example, if a patient seeks care in-hospital for a condition generally treated as outpatient, the inpatient data might not be captured. In addition, information recorded in visits to care providers outside of the EHR network may not transfer to the medical records attained by the study.

Claims data may lack treatment information for patients who are also enrolled in clinical trials, or could be missing codes for treatments or procedures that are not particularly costly or are available over the counter. [11,12] Data in existing registries may be missing due to different guidelines or attention to form completion, or missingness that was not resolved during study conduct. Out-of-network care and out-of-pocket payments may also be missing from claims. In some instances, patients are covered by multiple sources; for example, Medicare records for patients who have additional Health Maintenance Organization (HMO) coverage may be incomplete if the HMO covered costs. The high probability of missing diagnoses and treatments often necessitates excluding these patients, which in turn may impact both internal and external validity of study results.

Understanding the extent of missing data and the likely impact on the ability of the registry to meet its objectives is critical before determining whether to invest resources in linkage to a secondary data source. Creation of a robust set of specifications is the first step, allowing researchers to understand exactly what the EHR, claims database, or existing registry contains. Ideally, a feasibility assessment will be conducted in all or some of the data to scan the values within applicable fields for the number and percent of the data missing. This information can inform the decision as to whether the existing data source is sufficient to populate the registry, or if certain variables should be added to the registry CRF. Attaining and investigating a sample of data is also a good approach, as this may inform researchers about the potential for missing data as well as discrepancies in operational definitions of the variables. Where possible, medical codes (e.g., International Classification of Diseases [ICD9], Current Procedural Terminology [CPT], or National Drug Code [NDC]) are used to standardize data collection and avoid complications with interpreting text descriptions, which can vary significantly. However, care should be taken to understand what type of medical codes are being utilized. For example, procedure codes may vary in type, including CPT, EPIC, or Kellogg Cancer Center codes. A review of the codes contained within the sample data should be performed to understand if different coding systems are employed and whether they are easily and reliably distinguishable through a code type variable.

In addition, response options to variables of interest may also vary. For example, one site may report smoking status as 0: never, 1: former, 2: current, blank: missing while another site may report smoking status as 1: current, 2: former or never, 3: missing or unknown. Attaining and investigating a sample of data is critical to capturing these differences and ensuring valid data collection and integration. If EHR feeds will populate the registry prospectively, a sample or scan for missing data in retrospective fields can inform one's assessment of data to be collected in the future.

The processes for integrating secondary data sources with a patient registry are discussed elsewhere in the User's Guide (see Chapters 5 and 15-18).

3.2. Operational Strategies to Minimize Missing Data

While steps taken during the planning and design phases can reduce the likelihood of missing data, it is equally important to implement strategies to minimize missing data during the registry's operational phase. A plan for ongoing monitoring of the data is an essential tool for identifying and addressing missing data issues. As discussed in Chapter 11 of the User's Guide, a data management plan is critical for providing clear guidelines for review and handling of missing data. Queries may be issued to prompt sites to enter missing data or to correct issues with the data (for example, data values which are illogical, or out-of-range). Ideally, data review activities are conducted on an ongoing or periodic basis to avoid overburdening sites with a large number of queries to resolve at the conclusion of the study.

In addition to prompting sites to fill in missing data within the registry CRF, ongoing data review can identify trends in missing data or quality issues so that corrective actions can be taken before the conclusion of the study. For example, data review may reveal that the data collection guidelines are unclear for a specific data element, resulting in data not being entered into the system. Modifications to the data collection guidelines and additional training could be undertaken in this scenario to improve completion rates for that data element in the future. Alternately, the data review may identify issues with a particular site, such as high rates of missing data or loss to follow-up. [13] Additional one-on-one training may be needed to improve data quality and patient retention at that site.

4. Types of Missing Data

When considering the potential impact of the missing data on the registry findings, it is important to consider the underlying reasons for why the data are missing.[14] Missing data are typically grouped into three categories:

- Missing completely at random (MCAR). When data are MCAR, the fact that the data are missing is independent of the observed and unobserved data.[15] In other words, no systematic differences exist between participants with missing data and those with complete data. For example, some participants may have missing laboratory values because a batch of lab samples was processed improperly. In these instances, the missing data reduce the analyzable population of the study and consequently, the statistical power, but do not introduce bias: when data are MCAR, the data which remain can be considered a simple random sample of the full data set of interest. MCAR is generally regarded as a strong and often unrealistic assumption.

- Missing at random (MAR). When data are MAR, the fact that the data are missing is systematically related to the observed but not the unobserved data.[15] For example, a registry examining depression may encounter data that are MAR if male participants are less likely to complete a survey about depression severity than female participants. That is, if probability of completion of the survey is related to their sex (which is fully observed) but not the severity of their depression, then the data may be regarded as MAR. Complete case analyses, which are based on only observations for which all relevant data are present and no fields are missing, of a data set containing MAR data may or may not result in bias. If the complete case analysis is biased, however, proper accounting for the known factors (in the above example, sex) can produce unbiased results in analysis.

- Missing not at random (MNAR). When data are MNAR, the fact that the data are missing is systematically related to the unobserved data, that is, the missingness is related to events or factors which are not measured by the researcher. To extend the previous example, the depression registry may encounter data that are MNAR if participants with severe depression are more likely to refuse to complete the survey about depression severity. As with MAR data, complete case analysis of a data set containing MNAR data may or may not result in bias; if the complete case analysis is biased, however, the fact that the sources of missing data are themselves unmeasured means that (in general) this issue cannot be addressed in analysis and the estimate of effect will likely be biased.

While the complete case analysis of a dataset with MCAR data is unbiased, there is a common misperception that complete case analysis of a dataset with MNAR data will necessarily result in a biased estimate of effect. However, this is not so; in fact, whether missing data introduce bias into a complete case analysis depends on the causal structure of the missingness process. Details are given in Daniel et al.[7] and additional examples in Westreich[16]; but informally the complete case analysis will be unbiased due to missing data if the missingness is independent of the outcome under study, a condition that can be present whether the data are MAR or MNAR. However, if the missingness is not independent of outcome, it can be made so through analytic means only if the missingness is MAR. The import of the MAR vs. MNAR distinction is therefore not to indicate that there definitively will or will not be bias in a complete case analysis, but instead to indicate – if the complete case analysis is biased – whether that bias can be fully removed in analysis (see Section 5 for analytic strategies).

A comparison of the distribution of observed variables for patients with specific missing data to the distribution of those variables for patients for whom the same data are present can provide insights into the type of missing data occurring in the registry. More critical, however, is an appreciation of the qualitative expert knowledge and assumption that make up the causal structure of the variables, which are frequently encoded in a causal directed acyclic graph (DAG) [17,18]. Considerations to this effect can provide insights into the impact of the missing data as well as how to report and statistically manage it.

5. Analytic Implications and Management Strategies for Missing Data

5.1.1. Methods

Even when design and operational strategies are used to minimize the likelihood of missing data, missing data are likely to occur to some degree in patient registries and nonexperimental studies due to the fact that they do not dictate treatment or health care encounters but instead observe patient care as it occurs in routine practice. Additionally, while randomized clinical trials (RCTs) often limit their samples to a narrow and often healthier subset of patients, registries, by taking almost all comers, see a far more diverse patient sample. Some of these patients may be more likely to have missing data due to greater variation in language spoken, insurance coverage, education, general health, treatment history and other factors.

Analytic methods for handling missing data in clinical trials are well documented,[19,20] although these methods are not always applicable to observational studies. By their nature, clinical trials often have better collection of key variables, and randomization can provide statistical and analytic leverage for certain types of missing data. As an example, in observational studies, data can be missing, unknown or even not applicable for certain baseline characteristics. Dates can be missing or partially missing for disease diagnosis or progression, start or stop date of adverse events of interest, or treatment exposure. These challenges require careful consideration in choosing analytic methods.

As a first step in selecting an analytical strategy for handling missing data, it is important to understand the type, pattern and amount of missing data in the registry database. Statistical tools can be used to create graphical displays and frequency distributions for key variables, both overall and over time, which can help to identify trends in missing data. When evaluating the missing data, consideration should also be given to design or operational issues that may have affected the completeness of the data (e.g., revision of data collection guidelines to address common questions; changes during the life of the study, such as addition of a new data element or change in status from optional to required for a data element). To the extent possible, it is also important to understand the type(s) of missing data present in the database. As discussed previously, missing data can be categorized as MCAR, MAR, or MNAR. These designations influence the selection of the most appropriate analytical strategy for handling missing data, though as noted they do not necessarily indicate whether a complete case approach will be biased. Here it is important to note that MAR and MNAR cannot be formally tested, in ways which parallel how assumptions of no confounding and/or no uncontrolled confounding cannot be formally tested; it is thus highly recommended that sensitivity analyses be conducted to assess the robustness of the study results (see Section 6).

5.1.1. Complete Case Analysis

The complete case analysis strategy restricts the analysis to patients with complete data for all variables in the planned analysis; in other words, patients with missing data on any variable (including exposure, outcome, or included covariates) are excluded from analysis. This strategy is simple and is the default approach of many statistical software packages (and thus may happen without the investigator's intervention or even knowledge). Complete case analysis is inappropriate in many circumstances, however, as it implicitly assumes that the analysis properly accounts for missing data without regard for the underlying causal structure of the data generating mechanism, including the generation of missing values. Some insight can be obtained through consideration of a simple example: consider a longitudinal study which collects data at baseline, 30 days, and one year. Patients with missing data at one year may be missing for reasons unrelated to the outcome of interest (e.g., they changed their contact telephone number and did not inform study staff), or they may have missing data because they became too ill or died before completing the study. In the former case (changed telephone number) then a complete case analysis is likely to be unbiased; in the latter, deletion of these patients will bias the analysis by removing the sickest patients. Another way to think about the latter situation is that missingness was caused by the outcome (among other factors) and therefore the outcome is inexplicitly linked to the missing data. On the other hand, consider a study in which missingness in the outcome differs only by level of a confounder Z: individuals with Z=1 have a 20% chance of having a missing value for the outcome, while those with Z=0 have a 40% chance of missing outcome (and no other values are missing). Within strata of Z, missingness is effectively completely at random; thus, a model which controls for Z should be unbiased. From a causal DAG perspective, control for Z (the only cause of missingness) leads to independence of missingness and the outcome; and thus is sufficient for estimation of an unbiased effect.

When the data are MCAR, complete case analysis produces unbiased estimates; but as we note above, MCAR is not necessary for an unbiased complete case analysis. In either case, the precision of the estimates from a complete case analysis will be reduced due to loss of patients. In our above example, there is complete information on Z as well as the exposure; complete case analysis ignores that information. As a consequence, complete case analysis is perhaps most appropriate when the proportion of patients with missing data is relatively small (<10%); however in typical use, the conditions noted above (and described in detail by Daniel et al.) are not explained or explored. [21]

5.1.2. Single and Multiple Imputation

Unlike complete case analysis, patients with missing data are retained in the analysis when imputation methods are used. Imputation methods replace missing observations with values predicted in some manner, often from a model. In single imputation, the missing observation may be replaced with the sample mean or median (not recommended), with a predicted value of the variable (e.g., from a regression model, bootstrap, or a random dataset from multiple imputation), or with the value from a study patient who matches the patient with the missing data on a set of selected covariates. Another common form of single imputation is carry-forward in longitudinal data: if a patient has an observed lab value at time one, and is missing that lab value at time two, it is assumed that the value at time two is equal to the value at time one: the time one value is carried forward. Carry-forward options include last observation carried forward (LOCF), worst observation carried forward, or best observation carried forward.[22] Single imputation methods may introduce bias, sometimes substantial, if the data are not MCAR. Indeed, some single imputation approaches may introduce bias even if data are MCAR: consider a situation in which age is a confounder and missing completely at random for a large number of values, and in which we impute the median observed age as the value of all missing ages. This would result in ages which are incorrect for at least some individuals, and thus introduces information bias into the age variable. This, in turn, may well lead to incomplete adjustment for confounding by age (in a final regression model using imputed values) and thus a biased final effect estimate.

Single imputation based on a rich predictive model of sufficient covariates may be less prone to biases than median-based single imputation as an example, but single imputation will in any case be anti-conservative with respect to precision of effect estimates: imputing a single value does not properly acknowledge the uncertainty inherent in filling in missing data with their presumed values. For these reasons, single imputation is generally not recommended. When single imputation must be employed (for whatever reason, including potentially for reasons of technical limitations) investigators must report potential biases due to imputation method used (as per the age example above) and the anti-conservative nature of the confidence intervals or p-value.

One case where single imputation may be more appropriate is when dates are partially missing. For example, if the day element of an adverse event (AE) date is missing and is not retrievable, consideration can be given to imputation of the missing day element on the middle (e.g., 15th) day of the month. This date would need to be constrained by underlying study issues (for example, the date must be after the study enrollment date, before the AE resolution date, and study discontinuation date). If treatment changes during this month, the relationship with the timing of the treatment change should be considered when discussing the appropriate imputation method. This issue is closely related to that of interval censoring.

In general however, use of multiple imputation methods is strongly preferred to single imputation. In multiple imputation, multiple data sets are produced with different values imputed for each missing variable per data set, thus reflecting the uncertainty around the true values of the missing variables. The multiple values may be derived from the posterior probability distributions for the missing values.[15] As noted, the result of a multiple imputation process is multiple complete data sets for analysis from which a single summary finding is estimated. Standard errors are obtained through a combination of the between-model variance and the within-model standard errors, using Rubin's Rules for Imputation.[14] In practice, multiple imputation proceeds from a rich model for the missingness process; as such multiple imputation is very unlikely to result in biased estimates when data are MCAR. If data are MAR, multiple imputation will generally produce unbiased results if the model includes the correct set of covariates, and unlike single imputation will propagate error correctly. In the presence of MNAR data, multiple imputation in general cannot fully correct any bias due to missing data.

5.1.3. Inverse Probability Weighting

Inverse probability weights (IPW) are closely related to survey sampling weights[23] generalized to multiple complex variables and typically estimated with a parametric (and usually logistic) model although alternatives may be preferable.[24]

Broadly, inverse probability of missingness weights allow fully observed individuals to "stand-in" for partially observed individuals. For example, consider a study of 500 women and 500 men, in which 100 women and 200 men were missing a measurement of outcome. We could allow the 400 women to represent all 500 women in the study by giving each woman a weight of 1/(400/500) = 500/400 = 1.25. This number is obtained, note, by first estimating the probability that a woman is not missing: (400/500) and then taking the inverse of that number (500/400). Similarly, we could allow the 300 men to stand in for all 500 men by giving each one a weight of (300/500)-1 = 5/3 = 1.67. Once inverse probability of missingness weights are applied to the observed data, the reweighted data can be analyzed only among the complete cases; if modeling assumptions are met and the covariates considered fully account for joint predictors of missingness and the outcome the resulting analysis will be unbiased by missing data.

Like multiple imputation approaches, IPW requires data are MAR (or MCAR) to work; if data are MNAR, then modeling will fail to correct all bias from missing values (although it may reduce bias). IPW approaches to missing data (including right-censored data[25,26]) have the advantage of producing a unified approach to both confounding and missing data.

5.1.4. Maximum Likelihood Methods

Maximum likelihood estimation (MLE) is an analytic maximization procedure which provides the values of the model parameters that maximize the sample likelihood, i.e., the values that make the observed data "most probable". MLE has the advantage of using all available data, and does not require data to be sorted by a fixed number of study visits. Under the assumption of MAR, MLE is efficient and provides unbiased estimates. Calculating MLE's and fitting them into regression models for statistical inference often requires specialized software, especially when data are missing for predictor variables. This is still a challenge today, but as time goes by, more statistical packages are upgrading to contain MLE analysis capability. When data are missing for dependent variables only, likelihood based methods including the well-known mixed models for repeated measurements can be used for analyzing data with monotone or non-monotone missingness patterns. [27]

In observational research, not all studies have meaningful “visits”, or have data collected at a given visit for all subjects. For example, it is quite common that a subject can have unscheduled visits, and a lab test can be “missing” simply because the test was not ordered by the physician. Longitudinal data of this nature can be analyzed by the random intercepts model.[28] In this model, each subject is assumed to have a random effect, which follows a normal distribution. The time variable can be modeled as a random effect, a fixed effect, or both. [29] The model has the flexibility of allowing linear, quadratic or other forms of the time effect, and inclusion of the interaction effects between other covariates and time as fixed effects. If the repeated-measures are for a binary dependent variable, or count data, the above-mentioned benefits can be obtained by fitting the generalized liner mixed effects model. The generalized linear mixed model also assumes MAR.

6. Considerations for Reporting Findings from Studies with Missing Data

6.1. Reporting Guidelines

Missing data is common in patient registries and, depending on the extent and type of missing data, may affect the interpretation of results. As such, documenting how missing data were addressed when reporting registry findings is important in order to provide transparency and to allow readers to accurately interpret registry findings. To this end, two useful guidelines are the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement and the Patient-Centered Outcomes Research Institute (PCORI) Methodology Report.

The STROBE statement consists of a checklist of 22 items to address in reports of observational studies and includes missing data as one of the essential items to document. In the accompanying explanation, the STROBE authors provide the following guidance:

“We advise that authors report the number of missing values for each variable of interest (exposures, outcomes, confounders) and for each step in the analysis. Authors should give reasons for missing values if possible, and indicate how many individuals were excluded because of missing data when describing the flow of participants through the study. For analyses that account for missing data, authors should describe the nature of the analysis (e.g., multiple imputation) and the assumptions that were made (e.g., missing at random).” [30]

The PCORI Methodology Report, which describes standards for the conduct of patient-centered outcomes research, places a particular emphasis on missing data, with an entire section and five standards devoted to the topic. While the STROBE Statement focuses on items to include in reports of study findings, the PCORI standards address prevention of missing data, analytic approaches for addressing missing data, and reporting findings from studies with missing data. In addition to the items covered by the STROBE Statement, the PCORI standards include a requirement that investigators should consider the potential for missing data when developing a study protocol and plan for appropriate steps to minimizing the likelihood of missing data. Expected rates of missing data should also be set at the study outset, with comparisons made to actual rates of missing data during the analysis phase. Conducting sensitivity analyses (see Section 6.2 below) are also considered a mandatory component of study analysis and reporting in the PCORI standards, as are comparisons of the baseline characteristics of patients with or without missing data. In terms of reporting results, the PCORI standards require the information included in the STROBE Statement, plus a discussion of the potential impact of both the extent of missing data and the approach used to address missing data and incorporation of this information into the interpretation of the study findings. [31]

6.2. Recommended Sensitivity Analyses

In addition to the above approaches, “scenario-based” sensitivity analyses should be considered for missing data. Investigators can identify “worst case” scenarios for the missing data: for missing outcomes, one such “worst case” scenario might be to assume that all exposed missing outcomes are events, while all unexposed missing outcomes are nonevents (or vice-versa). Such scenario-based approaches can help set boundaries on causal effect size in ways that are useful for contextualizing main results. However, since scenario-based analyses are by their nature specific to the data and situation under study, it is important to consider carefully what questions are of most substantive relevance to the study question at hand. Ideally, sensitivity analyses using different analytic approaches for missing data should be pre-specified in the protocol or a separate data analysis plan, and not done post-hoc.

6.3. Missing Potential Outcomes and Causal Inference

To put the issue of missing data into perspective, it is useful to remember that, from the perspective of potential outcomes, causal effects can be defined as the expectation (over a population) in a contrast in individual potential outcomes – for example, the average risk of an outcome if the entire population had been exposed, contrasted with the average risk of an outcome if the entire population had been unexposed.[18] The central problem of causal inference is that it is not possible to observe more than a single potential outcome for any individual under study: that is, it is possible to observe what happens when an individual is exposed to X=1, but not to X=0 (or to any other value of X). As such, the central problem of causal inference is a problem of missing potential outcomes –a missing data problem.

Thus, there are numerous parallels between techniques for missing data and techniques for “regular” regression analysis to deal with confounding, including close parallels between assumptions like MCAR/MAR/MNAR and confounding.

7. Conclusions

In summary, missing data are a common area of concern for patient registries. While steps can and should be taken in the planning and operational phases of a registry to minimize the extent of missing data, the observational nature of registries makes it highly likely that at least some data will be missing from a registry database. Therefore, approaches to addressing missing data should be considered as part of the registry protocol and analysis plan development, and during the analytic phase. When selecting an approach, it is essential to understand the types of and reasons for missing data in order to select the most appropriate option. This paper describes several approaches currently used for analysis of observational datasets with missing data; however, statistical methods are advancing and evolving in this area, and new methods may be introduced in the future.

Once analyses are complete, any reports of registry findings should document the extent of missing data, explain how missing data were addressed, and consider the potential impact of missing data on the findings. This level of transparency is important to allow audiences to appropriately

REFERENCES &

1. Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. http://www.effectivehealthcare.ahrq.gov/registries-guide-3.cfm.
2. Mendelsohn AB, Dreyer NA, Mattox PW, et al. Characterization of Missing Data in Clinical Registry Studies. Therapeutic Innovation & Regulatory Science. 2015;49(1):146-54.
3. Applebaum, Katie M., Elizabeth J. Malloy, and Ellen A. Eisen. "Left Truncation, Susceptibility, and Bias in Occupational Cohort Studies." Epidemiology (Cambridge, Mass.) 22.4 (2011): 599–606. PMC. Web. 16 Sept. 2015.
4. Howards PP, Hertz-Picciotto I, Poole C. Conditions for bias from differential left truncation. Am J Epidemiol 2007;165:444–52.
5. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. 2003;158(9):915–920.
6. Cain KC, Harlow SD, Little RJ, et al. Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. Am J Epidemiol. 2011 May 1;173(9):1078-84. PMID: 21422059. PMCID: PMC3121224. Epub 2011/03/23.
7. Daniel, R.M., Kenward, M.G., Cousens, S.N., De Stavola, B.L. Using causal diagrams to guide analysis in missing data problems. Stat Methods Med Res. 2012;21:243–256.
8. Hernandez-Diaz S, Chambers C, Ephross S, et al. "Pregnancy Registries." In: Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. http://www.effectivehealthcare.ahrq.gov/registries-guide-3.cfm.
9. Lau, Bryan, Stephen R. Cole, and Stephen J. Gange. "Competing Risk Regression Models for Epidemiologic Data." American Journal of Epidemiology 170.2 (2009): 244–256. PMC. Web. 16 Sept. 2015.
10. Hripcsak, George, and David J Albers. "Next-Generation Phenotyping of Electronic Health Records." Journal of the American Medical Informatics Association : JAMIA 20.1 (2013): 117–121. PMC. Web. 16 Sept. 2015.
11. Meyer AM, Carpenter WR, Abernethy AP, et al. Data for cancer comparative effectiveness research: past, present, and future potential. Cancer. 2012 Nov 1;118(21):5186-97. PMID: 22517505. PMCID: PMC3431434. Epub 2012/04/21.
12. Meyer, AM, Liu H, Mack C, Carpenter WE, Brookhart MA. Improving the validity of CER through Principled Exploration of Data. Poster Presentation. 29th ICPE: International Conference on Pharmacoepidemiology & Therapeutic Risk Management, Montreal, Canada, August 15-28, 2013.
13. Cronenwett J, Leavy MB. "Case Example 25: Using Audits to Monitor Data Quality." In: Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. http://www.effectivehealthcare.ahrq.gov/registries-guide-3.cfm.

14. Rubin D. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, LTD; 1987.
15. Little Roderick JA, Rubin Donald B. Statistical Analysis with Missing Data. New York: Wiley; 1987.
16. Westreich, Daniel. “Berkson’s Bias, Selection Bias, and Missing Data.” Epidemiology (Cambridge, Mass.) 23.1 (2012): 159–164. PMC. Web. 16 Sept. 2015.
17. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10(1):37-48.
18. Hernan MA, Robins JM (2015). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming.
19. O'Kelly M, Ratitch B. Clinical Trials with Missing Data: A Guide for Practitioners. West Sussex, UK: John Wiley & Sons, LTD; 2014.
20. Dziura JD, Post LA, Zhao Q, et al. Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med. 2013 Sep;86(3):343-58. PMID: 24058309. PMCID: PMC3767219. Epub 2013/09/24.
21. Daniel, R.M., Kenward, M.G., Cousens, S.N., and De Stavola, B.L. Using causal diagrams to guide analysis in missing data problems. Statistical Methods in Medical Research, 21(3):243–256, 2012
22. Chang, Mark. "Missing Data Imputation and Analysis." Modern Issues and Methods in Biostatistics. Statistics for Biology and Health: Springer New York, 2011. 117-43.
23. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Amer Statist Assoc. 1952;47(260):663–685.
24. Rochat, Tamsen J. et al. “Detection of Antenatal Depression in Rural HIV-Affected Populations with Short and Ultrashort Versions of the Edinburgh Postnatal Depression Scale (EPDS).” Archives of Women’s Mental Health 16.5 (2013): 401–410. PMC. Web. 16 Sept. 2015.
25. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000;11:561–570.
26. Cole, Stephen R., et al. "Effect of Highly Active Antiretroviral Therapy on Time to Acquired Immunodeficiency Syndrome or Death Using Marginal Structural Models." American Journal of Epidemiology 158.7 (2003): 687-94.
27. O'Kelly, Michael, et al. "A Guide to Planning for Missing Data." Clinical Trials with Missing Data. John Wiley & Sons, Ltd, 2014. 71-129.
28. Allison PD, Handling Missing Data by Maximum Likelihood, paper 312-2012, SAS Global Forum 2012, http://support.sas.com/resources/papers/proceedings12/312-2012.pdf, accessed on 29 June 2015.
29. Littell RC, Milliken GA, Stroup WW, et al, SAS® for Mixed Models, Second Edition, SAS Institute, 2006
30. Oeyen S, Vandijck D, Benoit D, Decruyenaere J, Annemans L, Hoste E. Long-term outcome after acute kidney injury in critically-ill patients. Acta Clin Belg. 2007;62:337–40.
31. PCORI (Patient-Centered Outcomes Research Institute) Methodology Committee. 2013. “The PCORI Methodology Report.” pcori.org/research-we-support/research-methodology-standards